ISSN: 2581-3404 (Online) IF: 5.86 (SJIF)
IJIRTM, Volume-10, Special Issue-2, April-2026

Multimodal Deepfake Detection Using Visual, Audio, and Synchronization-Based Features

Yuvraj Sharma[1], Sanyam Malhotra[2], Chirag Sharma[3], Akshat Chawla[4], Vijay Bharti[5]

Department of Computer Science & Engineering (AI & DS), Panipat Institute of Engineering and Technology, Panipat, India [1,2,3,4,5]

yuvraj7124000@gmail.com[1], sanyammalhotra190@gmail.com[2], itschiragsharma077@gmail.com[3], akshat384chawla@gmail.com[4], Vijaybharti.cse-aids@piet.co.in[5]

Abstract. *Deepfake technology has caused revolutionary change in how someone can create along with manipulated digital media. Deep learning has made it so almost too easy to create any image, audio, and video that will feel completely real. Sometimes they are convincing enough that you can't tell the difference no matter how much you try to. This is everywhere on social media right now that spotting any fakes keeps getting harder. This presents a real problem for anyone who has any concerns about the truth online or their security. Due to all of this happening, we need to find better ways to detect deepfakes. In this paper, we take a close look at deepfake detection methods which are built on deep learning to combat the Deepfake methods. In here, we are going to break down the current approaches into four main groups which are image-based, video-based, audio-based, together with those combined with different media types.*

Keywords: Hand Gesture Recognition, virtual mouse, Finger Movement Detection, AI-Based Input System , Image Processing , Hand Landmark Detection.

Introduction

Digital platforms are known to completely change the way things are shared by users on social media now. In todays time, it feels like all the us are posting pictures, videos, and audio clips nonstop. Yes, it is helping us to stay in touch with each other but there's a massive downside. It's made it so much easier for anyone to twist or fake what is seen and heard [1][3]. Deepfakes tend to highlight this problem. They use deep learning tricks such as GANs as well as autoencoders to create fake audio and video that can fool just about anyone due to them looking so convincing [1][4]. With deepfake tools, you can swap someone's face in a video, mimic their voice, or even invent entire scenes with people who were never even there [1]. The real problem for anyone who is trying to spot deepfakes is how good they've gotten now [2][4]. Old-school forensic tricks just can't keep up them. Modern deepfakes tend to erase the obvious glitches, along with the flaws that are shown are also difficult to catch [4]. Early attempts at detection are usually focused on single images by using convolutional neural networks to hunt for oddities in pixels [7]. Thought they don't focus on the audio aspect or the lipsync with the person. So, researchers now have changed their ways. Now, they're looking at both video together with sound [2][5].

Models that are video based approach how things change within frames. They pick up the odd timings or movements in the given video [7]. Then comes the Audio-based methods. They turn tune into speech and then listen for clues which indicate if something's off [3]. The latest technology is to mix both of the visuals as well as the audio and how well they line up and work with each other rather than independently [6].

2. Background and Fundamental Concepts

With AI advancing so quickly, deepfakes have gotten good. Spotting them isn't just hard anymore, it also takes certain tricks [1][4].

A. Deepfake Generation Techniques

Now, deepfakes usually use structures for creating the fakes. Deep learning models like GANs and autoencoders are some of the common ones [4]. The generator keeps getting better with practice until eventually the fakes made by it look almost real. Autoencoders on the other hand take a different approach. They actually break down and reconstruct faces [1].Deepfake generation techniques can generally be categorized into several types. Face swapping which involves replacing the face of one individual with another while preserving the original facial movements [9]. Expression manipulation though, now this focuses on modifying a person's facial expressions to imitate another person's emotional behavior [9]. The Attribute editing techniques that are used to modify those specific and minute facial characteristics such as age, gender, or other visual features are quite advance [4]. All of these approaches are typically trained on large-scale datasets that in the end require substantial computational resources if we wish to produce highly realistic outputs [10].

B. Deepfake Detection As A Learning Problem

Combating deepfakes normally means training a model to tell some real stuff from fake stuff [1][2]. Deep learning detection methods focus on finding odd details that give fakes away. These include visual distortions in facial regions, pixel-level inconsistencies caused by imperfect image synthesis. Another method included is compression anomalies generated during video processing [9]. Such artifacts have often appeared around facial boundaries, lighting transitions, or texture patterns. However, as deepfake generation techniques improve, many of these visual cues became less noticeable, making detection increasingly difficult over time [1][4].

C. Deep Learning Architectures For Detection

Several deep learning architectures have been widely adopted for deepfake detection. It is because of their ability to extract complex patterns from visual data [2][6]. Convolutional Neural Networks (CNNs) are commonly used to extract spatial features from images along with individual video frames [2][11]. These architectures are particularly effective at identifying small visual artifacts, texture inconsistencies, and blending of irregularities that may appear in any manipulated media. Also, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are often used to capture different temporal dependencies in video frames [7]. By checking different sequences of frames, these systems can detect wrong patterns or inconsistencies. Quite recently a transformer-based architecture seem to have gained attention in deepfake detection research. Apparently it is related to their attention mechanisms which somehow allows them to model long-range dependencies and complex feature relationships [6]. Then there

are Hybrid systems that can combine multiple architectures. The have also been proposed so that we can leverage both spatial together with temporal information [6][7].

D. Modalities In Deepfake Detection

While Deepfake detection approaches may be categorized by using the type of input data they normally analyze, it is Image-based detections that focuses on individual frames. This attempt is further used to identify visual artifacts or inconsistencies present in a single image [2][11]. Then the Video-based detection methods are present which analyze the given temporal information across multiple frames. That data is used for detecting unnatural facial movements and any out of place motion patterns [7]. After that, Audio-based detection techniques examine different speech patterens and all the unique features to identify synthetic or manipulated noise signals [3]. Many Multimodal detection methods have been developed that have successfully combined all the multiple data modalities, such as visual as well as audio signals, which has finally provided us with improve detection accuracy and robustness [6]. Multimodal approach is getting a lot of attention. It has mostly to do with their being able to combine several types of data which ends up making the system more accurate.

3. Related Work

To be honest, deep learning is changing how the deepfake detections is done. Especially multimodal setups which can even combine different data types [1][2]. When it all started, all we had to focus on was spotting visual problems. Things like compression marks or unnatural patterns by using convolutional neural networks (CNNs) [2][11]. Now we have models such as ResNet, VGG, along with Xception which have become frequently used. That is simply due to them being really good at pointing at those tiny details that will reveal a fake [11]. When you enter this field, you will notice that there seem to be two main approaches. One is the old-school machine learning route and the other is deep learning approach. CNNs is a classic to use now because they are better at getting important features. The models can learn how to spot the differences present in a real and edited image. They do this by training on datasets that would cover both [4][8]. This not only makes the models perform better but also helps in reducing how much time and computing power you would need to get some good results [2]. Multimodal processes have started bring in excessive use. The structure here combines information from different yet at the same time similar sources. Those being spatial, temporal and spectral features [6]. Of course there are still challenges remain. Many models frequently don't work well outside the very limited datasets they were trained on [5]. Video quality, compression, lighting changes are just a few things that can throw systems in disarray. Combine that with deep fakes getting advance and the detection models have a heck of a time keeping up as well [4][9].

In short the field is moving from simply looking at pictures to using all types of signals which are images, video sequences, and audio. The next important thing is building a detection systems that can handle deepfakes in all shapes and forms.

4. Deepfake Detection Techniques

This section here is going to break it down current deepfake detection approach. It will explain how they work and what the kind of input data and the learning style they use.

A. Image-Based Detection Techniques

Deep fakes are known for targeting the image-based detection first. This is because to them focusing on analyzation of every single frame from a video is important. Especially the CNNs. The fake that has many

small inconsistencies that are analyzed by detection models based on CNNs [2]. These include but are limited to blurred or facial boundaries. They then result in imperfect face swapping, irregular lighting conditions that simply won't match the given environment [11]. There are some common choices for these models like VGGNet, ResNet, and Xception. It is their job to shift through bundles of real along with fake images and then learn all the subtle differences. But here comes the next issue. Looking at just the still frames would still certainly mean you could miss glitchy movements that tend to occur between frames. Plus, as fake generation algorithms keep getting better and better these visual defects get harder to spot.

B. Video-Based Detection Techniques

Considering the limitations of the image-based detection, we have video-based deepfake detection techniques. They focus on analyzing temporal information across multiple video frames.Instead of evaluating individual frames one by one, this approach examines the sequence in facial movements and behavioral patterns over time. Given time they are able to identify various problems that could indicate manipulation.Many deep learning approaches have been proposed for this purpose. Hybrid models that would combine Convolutional Neural Networks with Recurrent Neural Networks are commonly used to extract spatial features from individual frames. After that their temporal relationships temporal relationship is analyzed.With the incorporation of temporal information, video-based detection approaches are often more capable of identifying deepfake manipulations. However, these methods typically require a much greater computational power and larger training video datasets for effective training.

C. Audio-Based Detection Techniques

Audio-based deepfake detection focuses on identifying inconsistencies in voice signals generated by synthetic speech models. Given the various advancements in text-to-speech systems, detecting fake audio has become an important component of deepfake detection. It analyzes various characteristics of speech signals in order to identify synthetic or manipulated audio. These methods examine spectral features which are derived from speech signals, frequency patterns that may indicate artificial generation, and speech dynamics such as tone, rhythm, as well as articulation patterns. Deep learning architectures which are trained on spectrogram representations of audio signals could effectively distinguish between natural and artificially generated noises. Models often use spectrogram-based CNNs and architectures built for catching spoofed audio. But again using audio alone is not enough. High-quality voice can be faked, which is why audio should always work in tandem with other signals to achieve maximum affect.

5. Deep Learning Architectures and Datasets

CNNs play an important role in deepfake detection. It is in response to their inherent ability which allows them to identify small visual inconsistencies in wrong media. The systems are highly effective at detection of irregular textures, facial blending and inconsistent lighting patterns. All of which might occur during the deepfake generation process. Among the various CNN architectures there are a few that seem to be a cut above others.

A. Temporal and Hybrid Models

Hybrid models mean mixing of CNNs with models. It allows us to digest sequences which are RNNs and LSTMs. It deals with the odd blinking, jerking faces, and breaks during motion. While they perform better, they also seem to take more resources while also needing big and labeled datasets for identification.

B. Efficient And Scalable Architectures

Efficiency and scalability are two of the most important considerations given when something like deepfake detection systems is introduced. EfficientNet has emerged as a very promising architecture. It's use in compound scaling processes for balancing model depth, width, and resolution is exemplary. This approach is useful in allowing the model to maintain high accuracy and at the same time using less computational resources. EfficientNet-based models are certainly more useful in performing global frame-level analysis and for learning generalized features as compared to some other models out there.

C. Audio-Based Deep Learning Models

The Deep learning models can also be used for detecting manipulated audio in any deepfake content. These systems then often analyze different speech signals using spectrogram representations. The spectrogram representations can then capture both frequencies together. When combined with temporal characteristics of the audio data is can allow architectures such as spectrogram-based CNNs and models like AASIST to focus on identifying certain things. Different frequency-domain features, multiple temporal speech patterns, and acoustic inconsistencies are some of the main identifiers. While audio-based detection methods may effectively identify some voice manipulation, they are usually more effective when combined with visual detection approaches in multimodal systems.

D. Benchmark Datasets

Deepfake detection architectures is extremely depended on the datasets which are used for training and then evaluation. FaceForensics++ is one of the most commonly used datasets. It contains both real as well as fake videos which are generated with the help of multiple deepfake techniques [9]. The Deepfake Detection Challenge (DFDC) dataset is another that provides us with a large-scale collection of videos. This feature diversifies the actors and manipulating methods which ends up making it extremely useful any evaluation of detection models under different conditions [10]. Similarly in use is another dataset known as Celeb-DF. This one was designed to produce more realistic deepfake videos with as few as visual artifacts as possible. Although all those datasets have been contributed in deepfake detection research, they still end up containing limitations such as dataset bias, limited diversity, plus the lack of many real-world conditions [4]. Training models which are based on these benchmark datasets will end up allowing researchers to evaluate and then compare detection approaches used in a standardized condition. However, a single dataset then cannot fully represent the sheer complexity of real-world scenarios.

Table I: Comparison Of Deepfake Detection Approaches

Approach Type	Strengths	Limitations
Image-Based	Fast	No temporal info
Video-Based	Temporal analysis	High cost
Audio-Based	Speech detection	Weak alone
Multimodal	High accuracy	Complex

Following all these comparative analyses we can make many observations. Things like CNN-based models seem to remain a major part in deepfake detection research due to them possessing better feature extraction capabilities. On the other hand Video-based approaches result in enhancement of detection accuracy with the addition of temporal information. Although it does result in increases system complexity and computational cost. All the while multimodal approaches can be used to combine visual plus audio signals. This in turn provides the best detection performance, although it is still under active development and would require further optimization.

6. Comparative Analysis And Research Gaps

On the basis of the literature review and comparative analysis we have identified the following key research gaps [1][2][4]:

A. Lack Of Generalization

Generalization is difficult. Models which do well with familiar data tend to break when faced with new material. That's mostly due to training data not being wide enough.

B. Weak Multimodal Fusion Strategies

Multimodal fusion still requires a lot of work. Just combining modalities is a basic requirement, systems can't seem to be able to take all the advantages.

C. Limited Real-World Applicability

A large number of deepfake detection models are trained plus tested on some controlled datasets. The results indicate that they struggle against real-world conditions. The videos are affected by compression artifacts, background noise, and low-resolution recordings. All these factors work together in reducing the reliability of detection models in many practical applications.

D. High Computational Complexity

Top-performing models end up demanding too much of the computing power which tend to make them impractical for real-time use.

E. Insufficient Use Of Synchronization Features

There isn't enough focus given on synchronization features. Things such as checking if spoken words and lip movement really match are really underrated even though these cues can indicate fakery.

7. Future Directions

While looking through the research gaps made known in the previous section there seems to be many important directions for future work. One of the most important area is the development of more advanced multimodal fusion techniques. This will allow us to effectively combine information from different data sources such as visual, audio, and temporal signals [6]. Just improving the ability of detection models in generalizing of different datasets and manipulation techniques is research challenge. Further studies should also explore synchronization-based detection methods. As they could provide strong indicators of manipulated content [6].

A more promising direction is the design of lightweight deep learning models. These models can perform efficient real-time deepfake detection on devices with little and limited computational power. All this suggests that they may offer an improved feature representation and detection capabilities for deepfake analysis. By continuing research in these areas we will essentially be developing a more robust and scalable deepfake detection systems.

8. Conclusion

Deep fake are common but does not mean that the methos for their detections is lagging behind. Deep learning models are now programmed to solve the problems from all possible angles in the media such as images, video, audio or a mix.

Acknowledgements

We want to thank our mentor and our department's faculty and head of department for providing much needed guidance during the writing of this paper.

References:

[1] A. Heidari et al., "Deepfake Detection Using Deep Learning Methods: A Systematic Review," IEEE Access, 2024.
[2] M. Taeb and H. Chi, "Comparison of Deepfake Detection Techniques through Deep Learning," Journal of Artificial Intelligence Research, 2022.
[3] Y. Patel and M. Jain, "Deepfake Image Detection Using Machine Learning and Deep Learning," International Journal of Computer Science, 2024.
[4] S. Ahmed and D.-T. Dang-Nguyen, "Deepfake Detection: A Comparative Analysis," arXiv preprint arXiv:2308.03471, 2023.
[5] F. Mahmud et al., "Unmasking Deepfake Faces from Videos Using an Explainable Cost-Sensitive Deep Learning Approach," arXiv, 2023.
[6] A. H. Soudy et al., "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," Neural Computing and Applications, 2024.
[7] D. Karishma et al., "Deepfake Face Detection Using LSTM and CNN," International Journal of Intelligent Systems and Applications in Engineering, 2024.
[8] H. Nguyen and J. Yamagishi, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," IEEE, 2021.
[9] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision (ICCV), 2019.
[10] B. Dolhansky et al., "The DeepFake Detection Challenge Dataset," arXiv, 2020.
[11] M. Kumari et al., "Deepfake Detection Using XceptionNet," International Journal of Scientific Research in Science and Technology, 2025.